

## Pairwise Multiple Comparisons in SAS

Pairwise multiple comparisons are easy to compute using SAS Proc GLM. The basic statement is:

**means effects / options;**

Here, **means** is the statement initiator, **effects** indicates the treatment effects for which the means are desired and the **options** component allow for specification of the type of comparison. The set of possible options includes:

|                                      |                                |
|--------------------------------------|--------------------------------|
| ALPHA = p                            | Set the $\alpha$ level         |
| BON                                  | Bonferroni's method            |
| DUNCAN                               | Duncan's multiple range test   |
| DUNNETT <(formatted-control-value)>  | Dunnett's two-tailed test      |
| DUNNETTL <(formatted-control-value)> | Dunnett's lower-tailed test    |
| DUNNETTU <(formatted-control-value)> | Dunnett's upper-tailed test    |
| LSD                                  | Fisher's LSD                   |
| SCHEFFE                              | Scheffe's procedure            |
| SNK                                  | Student-Newman-Keuls procedure |
| TUKEY                                | Tukey's procedure              |
| WALLER                               | Waller-Duncan procedure        |

Using the Lettuce data, Fisher's LSD procedure is specified as follows:

```
options ls = 80;
Title "Analysis of the Lettuce Heads Data";
title2 "With LSDs for Pairwise Comparisons";
data Lettuce;
input Heads Fertilizer @@;
cards;
104    0  114    0  90    0  140    0
134   50  130   50 144   50  174   50
146  100  142  100 152  100  156  100
147  150  160  150 160  150  163  150
131  200  148  200 154  200  163  200
;
proc glm data = Lettuce;
  class Fertilizer;
  model Heads = Fertilizer;
  means Fertilizer / LSD;
run;
```

## Assessing Model Assumptions Using SAS

Residuals are key elements for assessing model assumptions. Residuals are constructed as follows:

$$\hat{e}_{it} = Y_{it} - \hat{\mu}_i.$$

Here,  $Y_{it}$  represents the response for the  $j^{\text{th}}$  observation on the  $i^{\text{th}}$  treatment and  $\hat{\mu}_i$  the estimated mean for the  $i^{\text{th}}$  treatment.

Assessing constant variance can be performed by plotting the residuals against either the treatment levels or the treatment means. In addition, Levene's test can be constructed using a little creative programming.

Assessing the assumption of normality can be performed by constructing a normal probability plot or running formal test of hypothesis, such as the Anderson-Darling test. Both of these approaches are implemented by SAS Proc Univariate.

As an example of these methods, consider the following experiment and data:

Example:

Seventy-eight (78) male workers were assigned at random to six different groups so that 13 were in each group. After training in a specific task, the pulse rate was measured for 20 seconds. Unfortunately, some (10) individuals withdrew from the experiment before their training was complete. The data from this experiment, along with some summary statistics are reproduced in the following table:

| Group |    |    |    |    |    |
|-------|----|----|----|----|----|
| 1     | 2  | 3  | 4  | 5  | 6  |
| 27    | 29 | 34 | 34 | 28 | 28 |
| 31    | 28 | 36 | 34 | 28 | 26 |
| 26    | 37 | 34 | 43 | 26 | 29 |
| 32    | 24 | 41 | 44 | 35 | 25 |
| 39    | 35 | 30 | 40 | 31 | 35 |
| 37    | 40 | 44 | 47 | 30 | 34 |
| 38    | 40 | 44 | 34 | 34 | 37 |
| 39    | 31 | 32 | 31 | 34 | 28 |
| 30    | 30 | 32 | 45 | 26 | 21 |
| 28    | 25 | 31 | 28 | 20 | 28 |
| 27    | 29 |    |    | 41 | 26 |
| 27    | 25 |    |    | 21 |    |
| 34    |    |    |    |    |    |

The SAS code for running the ANOVA and extracting the residuals using Proc GLM are provided below. Also included in the program code are the methods of using Proc Univariate to extract the normal probability plot and the Anderson-Darling test for normality. Lastly, a second Proc GLM analysis is performed to produce Levene's test for equality of the variances.

```

options pageno = 1;
title "Analysis of Pulse Rate for 6 Treatment Groups";
data task;
input pulse group @@;
cards;
27 1 29 2 34 3 34 4 28 5 28 6
31 1 28 2 36 3 34 4 28 5 26 6
26 1 37 2 34 3 43 4 26 5 29 6
32 1 24 2 41 3 44 4 35 5 25 6
39 1 35 2 30 3 40 4 31 5 35 6
37 1 40 2 44 3 47 4 30 5 34 6
38 1 40 2 44 3 34 4 34 5 37 6
39 1 31 2 32 3 31 4 34 5 28 6
30 1 30 2 32 3 45 4 26 5 21 6
28 1 25 2 31 3 28 4 20 5 28 6
27 1 29 2 41 5 26 6
27 1 25 2 21 5
34 1
;
proc print data = task;
run;
title2 "Analysis of Variance for Raw Data";
proc glm data = task;
class group;
model pulse = group;
output out = new r = residuals;
run;
title2 "Residual Plot for the Raw Data";
proc plot data = new;
plot residuals*group;
run;
data task; set task;
if group = 1 then z = abs(pulse - 31);
if group = 2 then z = abs(pulse - 29.5);
if group = 3 then z = abs(pulse - 34);
if group = 4 then z = abs(pulse - 37);
if group = 5 then z = abs(pulse - 29);
if group = 6 then z = abs(pulse - 28);
run;
title2 "Levenes Test for Equality of Variances - Raw Data";
title3 "Analysis of Variance for Median Based Residuals";
proc glm data = task;
class group;
model z = group;
run;
title2 "Assessing the Assumption of Normal Residuals";
title3 "Using the Normal Probability Plot and the Wilk's-Shapiro Test";
proc univariate data = new normal plot;
var residuals;
run;

```

The above SAS code can be found on the STAT 512 web page in a document titled "SAS\_Pulse\_Example.txt." To extract this code, open the document, copy the code to the clipboard and paste in the SAS editor.

## Analysis of a Completely Randomized Design with a Two-way Treatment Structure

### Example

An experiment was carried out to study the effects of two factors (Time of Bleeding: morning or afternoon, and Diethylstilbestrol: with or without) on plasma phospholipid in lambs. Five lambs were assigned at random to each of the four possible treatment groups and plasma levels of phospholipid subsequently measured. Unfortunately, two of the lambs died before the experiment could be completed. The data from this experiment are as follows:

|                         |          | Time of Bleeding |          | Marginal Means |
|-------------------------|----------|------------------|----------|----------------|
|                         |          | A.M.             | P.M.     |                |
| Diethyl-<br>stilbestrol | No       | 8.53             | 39.14    | 24.00          |
|                         |          | 12.53            | 26.20    |                |
|                         |          | 14.00            | 31.33    | or             |
|                         |          | 10.80            | 45.80    |                |
|                         |          | (11.465)         | 40.20    | 25.39          |
|                         | Yes      |                  | (36.534) |                |
|                         |          | 17.53            | 23.80    | 23.06          |
|                         |          | 21.07            | 28.87    |                |
|                         |          | 20.80            | 25.06    | or             |
|                         |          | 17.33            | 29.33    |                |
|                         | 20.07    | (26.765)         | 22.65    |                |
|                         | (19.360) |                  |          |                |
| Marginal<br>Means       | 15.41    | 31.65            | 23.53    |                |
|                         | or       | or               | or       |                |
|                         | 15.85    | 32.19            | 24.02    |                |

The above design is not balanced with respect to the sample sizes within each treatment combination. This poses a dilemma when it comes to computing the sum of squares for the two main effects (Diethylstilbestrol: Yes or No, and Time of Bleeding: A.M. or P.M.). To compute the sum of squares for these components requires the marginal means and the grand mean. However, there are at least two ways to compute the marginal means: average the cell means or average the response values across all levels of the other factor. For the grand mean you could average the marginal means or average the response for all observations. In the above example the marginal means and the grand mean have been computed using both approaches (the average of the means is shown on top and the average of the response values shown on the bottom). Now let's see what SAS does.

The following SAS code will input the data, produce the analysis of variance table, compute the cell, marginal and grand means.

```

options pageno = 1 ps = 40;
title 'Plasma Phospholipid Variation in Lambs';
title2 'CRD with Two-way Treatment Structure - Unequal Cell Sizes';
data a;
input hormone $ time $ Y @@;
cards;
  NO AM 8.53    NO AM 12.53    NO AM 14.00    NO AM 10.80
  NO PM 39.14   NO PM 26.20   NO PM 31.33   NO PM 45.80   NO PM 40.20
YES AM 17.53   YES AM 21.07   YES AM 20.80   YES AM 17.33   YES AM 20.07
YES PM 23.80   YES PM 28.87   YES PM 25.06   YES PM 29.33
;
proc print data = a;
run;
title3 'The Usual Analysis for a CRD with Two-way Treatment Structure';
title4 'With lsmeans for main effects (marginal means comparisons) and';
title5 'interaction effects (simple cell means comparison)';
proc glm data = a;
  class hormone time;
  model y = hormone time hormone*time;
  lsmeans hormone time hormone*time / pdiff stderr;
run;

```

### Assessing Model Assumptions

Before discussing the interpretation of the results from the analysis of variance, we should probably assess whether the assumptions of the model are valid. To do this requires the residuals. SAS Proc GLM will create a new data set with the residuals and means if requested. For the above analysis, inserting

```
output out = new p = means r = residuals;
```

before the last run statement will produce a new SAS data set called *new* which will contain the information contained in the data set *a* along with two new variables called *means* and *residuals*.

Following the SAS Proc GLM statements with

```
Proc Univariate data = new normal plot;
  var residuals;
run;
Proc Plot data = new;
  plot residuals*means;
run;
```

will produce the test of the normality assumption and the residuals plotted against the cell means for assessing constant variance. Of course we could perform Levene's test for equality of variances, but we will assume that the residual plot is sufficient for our purposes.

## Assessing the Interaction

Assuming that the assumptions were satisfied, the next step in the analysis is the interpretation of the interaction. From the SAS Proc GLM analysis, the interaction is found to be highly significant (Pvalue = 0.0011). This would indicate that the main effects are not interpretable (refer to the interaction plots found on pages 6, 7 and 8 of the chapter 6 notes). However, one should always look at the plot of the cell means to determine the form of the interaction and whether it will affect the interpretation of the main effects analysis (refer to the interaction plot on page 8 of the notes in chapter 6). To plot the cell means use the following SAS statements at the bottom of the code:

```
options ps = 40;  
proc plot data = new;  
    plot means*hormone = time;  
    plot means*time = hormone;  
run;
```

Now, you can interpret whether the interaction adversely affects the interpretation of the main effects.

If the interaction severely affects the interpretation of the main effects, the least squares means (**LSMEANS**) analysis can be used to assess differences between cell means (**simple effects**).

## Analysis of Experiments Involving Two Factors with One Replicate Per Treatment Combination

### Example

An experimental design was developed to compare yield (in pounds) for two varieties of wheat (factor A) and four fertility regimes (factor B). The experiment is composed of a two-way treatment structure (A crossed with B) in a completely randomized design structure. However, only a single replication is taken for each treatment combination. The data collected for this experiment are given in the following table:

| VARIETY | REGIME |      |      |      |
|---------|--------|------|------|------|
|         | 1      | 2    | 3    | 4    |
| 1       | 35.4   | 36.7 | 34.8 | 39.5 |
| 2       | 37.9   | 38.2 | 36.4 | 40.0 |

To analyze data it must be assumed that the interaction between Regime and Variety is negligible and therefore the interaction term can be used as the error for the model. The reasoning behind this approach is discussed in the class notes. Enter the following SAS code and submit for SAS processing:

```
options pageno = 1;
data wheat;
input variety regime yield @@;
cards;
1 1 35.4  1 2 36.7  1 3 34.8  1 4 39.5
2 1 37.9  2 2 38.2  2 3 36.4  2 4 40.0
;
proc print data = wheat;
run;
proc glm data = wheat;
class variety regime;
model yield = variety regime;
means variety regime / lines lsd;
run;
options pagesize = 40;
proc plot data = wheat;
plot yield*regime=variety;
plot yield*variety=regime;
run;
```